

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

PRÁCTICA 2

PRESENTADO POR: *ANDRES RICARDO SANABRIA GARAY*

Presentación

En esta práctica se trabaja un caso orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

Competencias

En este ejercicio se desarrollan las siguientes competencias del Máster de Data Science:

- ✓ Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- ✓ Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos

Los objetivos concretos de esta práctica son:

- ✓ Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- ✓ Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- ✓ Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- ✓ Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- ✓ Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.

- ✓ Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- ✓ Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Contenido

Descripción de la práctica a realizar

1. Descripción del dataset
2. Integración y selección
3. Limpieza de los datos
 - 3.1 ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.
 - 3.2 Identifica y gestiona los valores extremos.
4. Análisis de los datos.
 - 4.1 Selección de los grupos de datos que se quieren analizar/comparar (p. ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)
 - 4.2 Comprobación de la normalidad y homogeneidad de la varianza.
 - 4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.
5. Representación de los resultados
6. Resolución del problema
7. Código
8. Video

Descripción de la práctica a realizar

El objetivo de esta actividad será ejecutar tareas de integración, transformación, limpieza, validación y análisis del dataset “Heart Attack Analysis & Prediction dataset” disponible en Kaggle, en el siguiente enlace:

<https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>

Con este conjunto de datos y siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes: descripción del dataset, integración y selección, limpieza de los datos, análisis de los datos, representación de los resultados, resolución del problema, código y video explicativo.

Es importante mencionar que el código en R, con el cual se obtienen los resultados que se presentan en cada uno de los puntos de este documento, se encuentra en el archivo ***Practica_2.Rmd*** que se adjunta en el mismo repositorio de Github de esta práctica.

1. Descripción del dataset

Este conjunto de datos disponible en Kaggle fue cargado por Rashik Rahman, se compone de 14 variables y 303 observaciones y se pretende que sirva de referencia para el análisis y la predicción de ataques al corazón en seres humanos.

El conjunto de variables son las siguientes:

age:	Edad del paciente
sex:	Sexo del paciente (0 = Femenino, 1 = Masculino)
cp:	Tipo de dolor de pecho (1: Angina típica, 2: Angina atípica, 3: Dolor no anginal, 4: Asintomático)
trtbps:	Presión sanguínea en reposo (en mm Hg)
chol:	Colesterol en mg/dl obtenido mediante sensor BMI
fbs:	Glucemia en sangre > 120 mg/dl (1 = true; 0 = false)
restecg:	Resultados del electrocardiograma en reposo [0: normal, 1: teniendo ondas anormales ST-T (T inversiones de onda y/o ST elevación o depresión > 0.05 mV), 2: muestra probable o definitiva hipertrofia ventricular izquierda por el criterio de Estes]
thalachh:	Máximo ritmo cardiaco alcanzado
exng:	El ejercicio produce angina (1 = si, 0 = no)
oldpeak:	Depresión del ST inducida por el ejercicio en relación con el reposo
slp:	Pendiente del segmento ST de ejercicio máximo (0 = sin pendiente, 1 = plana, 2 = pendiente baja)
caa:	Número de vasos sanguíneos principales (0-3)
thall:	Talasemia (0 = nula, 1 = defecto fijo, 2 = normal, 3 = defecto reversible)
output:	0= menos chance de tener un ataque al corazón, 1= mayor chance de tener un ataque al corazón.

Según se describe en la plataforma, estos datos fueron obtenidos por el autor en la web mediante técnicas de crawling.

El objetivo será responder a partir de los datos suministrados a las pregunta ¿es posible predecir si una persona tiene mayor o menor chance de sufrir un ataque al corazón?, ¿con que precisión?

2. Integración y selección

Si bien como se mencionó en el punto anterior quién publicó los datos en Kaggle afirma que provienen de la Web y los obtuvo mediante técnicas de Crawling, una búsqueda más

detallada nos permitió encontrar que en el “Machine Learning Repository” de la University of California Irvine (UCI) disponible en:

<https://archive.ics.uci.edu/ml/datasets/heart+disease>

Se encuentran cuatro bases de datos de cuya integración se obtiene el dataset con el que se trabaja este ejercicio. De acuerdo con la información publicada en el repositorio de la Universidad los autores de estas bases son:

- A. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- B. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- C. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- D. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

El conjunto original contiene 76 atributos. Sin embargo, en todos los experimentos que se han publicado se ha utilizado un subconjunto de 14 atributos (los que se describieron en el primer punto). Los nombres y números de seguridad social de los pacientes fueron reemplazados para anonimizar la información.

Ahora bien, como parte del proceso de integración se pueden haber generado duplicados, más cuando dos de las bases de datos provienen del mismo país (Switzerland), por lo que con las funciones `duplicated()` y `unique()` en R se realiza la corrección pasando a 302 observaciones, ya que una se encontraba duplicada (ver código en archivo *Practica_2.Rmd*).

En este punto también se ha corregido la tipificación de cada variable, ya que para los análisis es un paso fundamental distinguir entre variables numéricas y categóricas (aún cuando éstas se representen mediante números).

Con relación al proceso de selección, podríamos obtener subconjuntos de datos por ejemplo de acuerdo con el sexo del paciente o su rango de edad. Esto lo haremos más adelante y no descartaremos trabajar con el conjunto de datos completo.

3. Limpieza de los datos

3.1. ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

En el conjunto de datos no se identifican valores vacíos o no disponibles. Sin embargo, varios atributos tienen ceros dentro de sus valores. En estos casos lo importante es conocer su significado dentro del atributo y cuando corresponde con un valor nulo o con un dato que tiene un significado diferente.

Para el conjunto de datos escogido en esta práctica la variable asociada con la Talasemia “thall”, define el valor de 0 como nulo (revisar en el primer punto en la descripción del dataset). Y como se muestra en el archivo que contiene el código “*Practica_2.Rmd*”, para este atributo hay dos valores nulos.

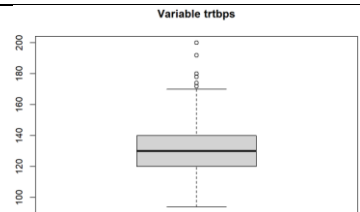
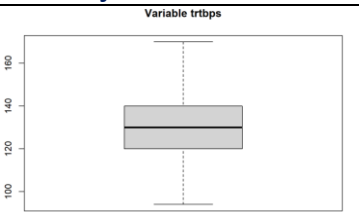
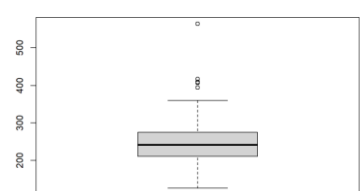
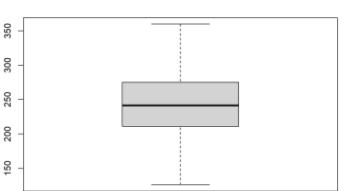
Ahora bien, ¿Qué debemos hacer con ellos?, la intuición nos lleva a pensar en el hecho de imputar valores a través de alguno de los métodos explicados en el material del curso. No obstante, debemos recordar que la Talesemia es una enfermedad es una enfermedad genética que afecta la producción de hemoglobina, causando anemia y otros problemas de salud, imputar un valor a un paciente no es adecuado en este caso y teniendo en cuenta que la cantidad de nulos es menor al 1% de las observaciones, se toma la decisión de prescindir de estos dos registros.

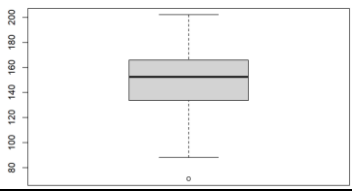
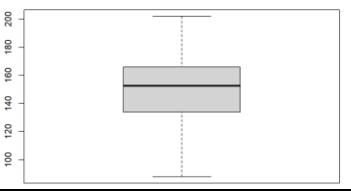
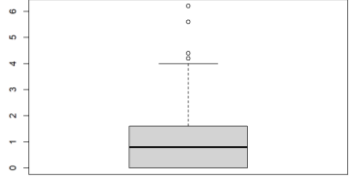
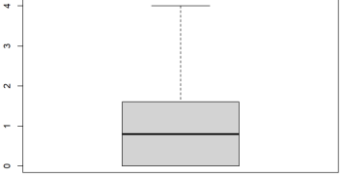
En este punto, el conjunto de datos ahora cuenta con 300 observaciones.

3.2. Identifica y gestiona los valores extremos.

La gestión de los valores extremos se realizó como se presenta de forma detallada en el archivo con el código “*Practica_2.Rmd*”. Básicamente las variables cuantitativas fueron objeto de análisis en busca de outliers, definidos para esta práctica como aquellos valores que se encuentran fuera del rango de $Q1 - 1.5 \cdot RI < x < Q3 + 1.5 \cdot RI$, donde x es el valor del atributo que estamos examinando, $Q1$ y $Q3$ corresponden a los cuartiles 1 y 3 respectivamente y RI es el rango intercuartílico que se obtiene de la diferencia entre $Q3$ y $Q1$. Con esta descripción se presenta la siguiente Tabla 1:

Tabla 1. Resumen del tratamiento aplicado a los outliers y el resultado

Variable	Outliers	Con ajuste de Outliers	Observaciones
trtbps			Los outliers se reemplazan con la mediana del atributo
chol			Se reemplazan los valores por el valor calculado de Q3

thalachh			Se reemplaza el outlier por el valor calculado de Q1
oldpeak			Debido a que no se trata de una distribución normal se utiliza un método de interpolación para reemplazar los outliers

Fuente: Elaboración propia. El código, las gráficas y una descripción más detallada se encuentra en el archivo *Practica_2.Rmd* disponible en la carpeta código del repositorio de GitHub de esta práctica

4. Análisis de los datos

4.1. Selección de los grupos de datos que se quieren analizar/comparar (p. ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)

Antes de proceder con la generación de grupos, realizaremos un poco de estadística descriptiva adicional a la que ya hemos examinado durante el desarrollo de la práctica. Esta es una actividad útil para los análisis y el trabajo que viene después, esta actividad se resume en la Tabla 2.

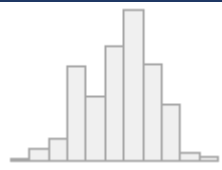
Data Frame Summary

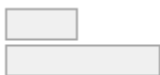
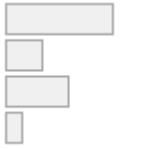
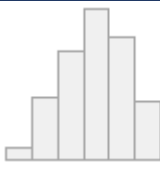
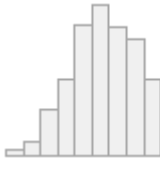
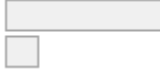

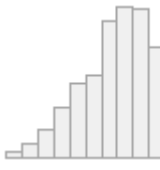
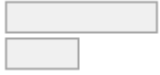
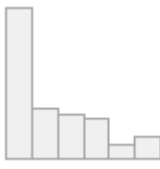
df_heart_attack



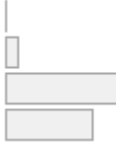

Dimensions: 300 x 14

Duplicates: 0

Tabla 2. Estadística descriptiva sobre los atributos del dataframe

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	age [integer]	Mean (sd) : 54.4 (9.1) min ≤ med ≤ max: 29 ≤ 56 ≤ 77 IQR (CV) : 13.2 (0.2)	41 distinct values		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
2	sex [factor]	1. 0 2. 1	95 (31.7%) 205 (68.3%)		0 (0.0%)
3	cp [factor]	1. 0 2. 1 3. 2 4. 3	142 (47.3%) 50 (16.7%) 85 (28.3%) 23 (7.7%)		0 (0.0%)
4	trtbps [numeric]	Mean (sd) : 130.1 (15.2) min ≤ med ≤ max: 94 ≤ 130 ≤ 170 IQR (CV) : 20 (0.1)	43 distinct values		0 (0.0%)
5	chol [numeric]	Mean (sd) : 244 (44.9) min ≤ med ≤ max: 126 ≤ 241.5 ≤ 360 IQR (CV) : 64 (0.2)	147 distinct values		0 (0.0%)
6	fbs [factor]	1. 0 2. 1	256 (85.3%) 44 (14.7%)		0 (0.0%)
7	restecg [factor]	1. 0 2. 1 3. 2	146 (48.7%) 150 (50.0%) 4 (1.3%)		0 (0.0%)
8	thalachh [numeric]	Mean (sd) : 149.9 (22.5) min ≤ med ≤ max: 88 ≤ 152.5 ≤ 202 IQR (CV) : 32.1 (0.1)	91 distinct values		0 (0.0%)
9	exng [factor]	1. 0 2. 1	202 (67.3%) 98 (32.7%)		0 (0.0%)
10	oldpeak [numeric]	Mean (sd) : 1 (1) min ≤ med ≤ max: 0 ≤ 0.8 ≤ 4 IQR (CV) : 1.6 (1.1)	38 distinct values		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
11	slp [factor]	1. 0 2. 1 3. 2	21 (7.0%) 139 (46.3%) 140 (46.7%)		0 (0.0%)
12	caa [factor]	1. 0 2. 1 3. 2 4. 3 5. 4	173 (57.7%) 65 (21.7%) 38 (12.7%) 20 (6.7%) 4 (1.3%)		0 (0.0%)
13	thall [factor]	1. 0 2. 1 3. 2 4. 3	0 (0.0%) 18 (6.0%) 165 (55.0%) 117 (39.0%)		0 (0.0%)
14	output [factor]	1. 0 2. 1	137 (45.7%) 163 (54.3%)		0 (0.0%)

Fuente: Generado con la librería [summarytools](#) 1.0.1 (R version 4.2.2)

En este punto, procedemos a generar los grupos a partir de nuestro conjunto de datos que nos gustaría analizar y comparar. Generaremos un grupo de hombres y otro de mujeres, así como un grupo para las personas menores de 55 años y otro para las de 55 años o más.

Los dos grupos obtenidos a partir del sexo tienen un tamaño de 95 mujeres y 205 hombres.

Los dos grupos obtenidos por edad corresponden a 141 personas menores de 55 años y 159 personas de 55 años o más.

Los detalles de como se obtuvieron estos grupos se pueden revisar en el archivo “*Practica_2.Rmd*”.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Las pruebas que se realizarán en este punto se aplicarán al conjunto completo de los datos, así como a cada uno de los grupos que se determinaron en el ítem anterior.

Con el objetivo de comprobar la normalidad de los datos, iniciamos aplicando los test de Kolmogorov-Smirnov y de Shapiro-Wilk. La hipótesis nula es que, el atributo bajo análisis tiene una distribución normal, en este caso si el p-valor es menor al nivel de significancia, normalmente igual a 0.05, entonces la hipótesis nula es rechazada y se concluye que los datos no cuentan con una distribución normal

La Tabla 3 presenta los resultados de normalidad que arrojan los dos test para el conjunto completo de datos

Tabla 3. Pruebas de normalidad sobre las variables cuantitativas

Variable	¿El atributo sigue una distribución normal?	
	Test de Kolmogorov-Smirnov	Test de Shapiro-Wilk
age	Si	Si
trtbps	No	No
chol	Si	Si
thalachh	Si	No
oldpeak	No	No

Fuente: Elaboración propia.

En este caso diremos que solo las variables age y chol siguen una distribución normal.

La Tabla 4 presenta los resultados de pruebas de normalidad para los grupos de hombres y mujeres.

Tabla 4. Pruebas de normalidad sobre las variables de los Grupos de mujeres y hombres

Variable	¿El atributo sigue una distribución normal para las mujeres?		¿El atributo sigue una distribución normal para los hombres?	
	Test de Kolmogorov-Smirnov	Test de Shapiro-Wilk	Test de Kolmogorov-Smirnov	Test de Shapiro-Wilk
age	Si	Si	Si	No
trtbps	Si	Si	No	No
chol	Si	Si	Si	Si
thalachh	Si	No	Si	No
oldpeak	No	No	No	No

Fuente: Elaboración propia.

La Tabla 5 presenta los resultados de pruebas de normalidad para los grupos juventud y experiencia.

Tabla 5. Pruebas de normalidad sobre las variables de los Grupos juventud y experiencia

Variable	¿El atributo sigue una distribución normal para los jóvenes?		¿El atributo sigue una distribución normal para los experimentados?	
	Test de Kolmogorov-Smirnov	Test de Shapiro-Wilk	Test de Kolmogorov-Smirnov	Test de Shapiro-Wilk
trtbps	Si	Si	Si	No
chol	Si	Si	Si	Si
thalachh	Si	No	Si	No
oldpeak	No	No	No	No

Fuente: Elaboración propia.

Para comprobar la homogeneidad de varianzas entre grupos lo haremos entre el grupo de hombres y mujeres empleando los test de Levene para los casos en los que la distribución es normal y Fligner Killen cuando los datos no cumplen la condición de normalidad.

En estos casos se asume que la hipótesis nula corresponde con la igualdad de varianzas en los diferentes grupos de datos, por lo que p-valores inferiores al nivel de significancia indicarán heterocedasticidad.

Los resultados, que se pueden ver de forma detallada en el archivo que contiene el código en R (el archivo *Practica_2.Rmd*), fueron:

Con el test de Levene se encontró Homocedasticidad entre el atributo age y chol.
Con el test de Fligner-Killeen se encontró homogeneidad de varianzas entre los atributos trtbps y thalachh, así como entre este último y oldpeak.

Esto es porque el valor de p es mayor que la significancia podemos afirmar que entre los atributos que se compararon hay homocedasticidad.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

En este punto se inicia comparando si existen diferencias estadísticamente significativas entre el grupo de hombres y mujeres con el atributo chol a través de la prueba t Student ya que este atributo tiene una distribución normal y presenta homocedasticidad. Así mismo, se utiliza la prueba de Wilcoxon para las variables trtbps, thalachh y oldpeak, pues están presentan una distribución diferente.

Como resultado de la comparación de los cuatro atributos se encuentra que entre el grupo de hombres y mujeres solo el atributo chol tiene diferencias estadísticamente significativas. Se repitió este mismo procedimiento para los grupos juventud y experiencia obteniendo que los cuatro atributos para estos dos grupos presentan diferencias estadísticamente significativas.

Esto nos lleva a pensar que la edad parece tener una fuerte influencia sobre los demás atributos dentro del conjunto de datos incluso más que el sexo. Pero vamos a aclarar esto con nuestra siguiente prueba estadística, un análisis de correlación.

Contraste de Hipótesis

Para el contraste de Hipótesis utilizamos en los atributos que tienen una distribución normal el test de t Student, mientras en los que no lo tienen la prueba de Wilcoxon-Mann-Whitney. La Hipótesis nula será que las medias de las muestras son iguales. Si el valor de p es menor

al nivel de significancia = 0.05, rechazaremos la hipótesis nula y diremos que las medias de las dos muestras son significativamente diferentes.

El resultado de esta actividad nos muestra que la media del atributo chol es significativamente diferente para las mujeres y los hombres, así como para los grupos juventud y experiencia, en los que además todos los atributos comparados son significativamente diferentes.

Finalmente, los atributos trtbps, thalachh y oldpeak, entre los grupos de hombres y mujeres no tienen diferencias significativas.

Análisis de correlación

El análisis de correlación entre las variables cuantitativas muestra que ninguno de los resultados se acerca a 1 o -1, ni siquiera superan 0.5 o están por debajo de -0.5, lo que nos permite suponer una débil correlación entre estas variables o dependencia, algo que será clave para la elaboración de nuestros modelos. Los resultados detallados se encuentran en el archivo *Practica_2.Rmd*

Modelo de regresión logística

Con la intención de alcanzar el objetivo de responder las preguntas descritas en el primer punto se plantea la generación de un modelo de regresión logística en el que la posibilidad de tener un mayor o menor chance de sufrir un ataque cardiaco será la variable dependiente e inicialmente todas las demás variables o atributos corresponderán con el conjunto de variables explicativas.

En este punto se vuelve a trabajar con el conjunto de datos completo que se ha venido limpiando (antes de la división en grupos de mujeres y hombres o jóvenes y experimentados) y se divide en un conjunto de entrenamiento y otro de prueba, en una proporción de 80 – 20 respectivamente.

Una vez generado el modelo se obtiene una precisión del 85%, con una sensibilidad del 79.41% y una especificidad del 92.31%.

5. Representación de los resultados

El archivo *Practica_2.Rmd* que está en R Markdown y su versión ejecutada *Practica_2.html* contienen todo el código del cual se han extraído los puntos aquí desarrollados, así como las gráficas empleadas.

6. Resolución del problema

En el inicio de esta práctica planteamos las preguntas:

¿Es posible predecir si una persona tiene mayor o menor chance de sufrir un ataque al corazón?, ¿con que precisión?

La respuesta es sí, con una precisión del 85%, sensibilidad del 79.41% y especificidad del 92.31%.

Lo anterior luego de generar un modelo de regresión logística en el que la variable objetivo fue output (mayor o menor chance de sufrir un ataque cardiaco). Es importante tener en cuenta que este modelo tiene mejores posibilidades de predicción correcta si del paciente bajo análisis se obtienen datos que se encuentran dentro de los valores límites del conjunto de datos que se utilizó para esta práctica.

Por último, también se debe mencionar que el dataframe preprocesado y al que se le realizaron las labores de limpieza fue guardado con el nombre `hear_clean.csv` en la carpeta de la práctica en GitHub.

7. Código

El código de esta práctica se encuentra en el archivo *Practica_2.Rmd* que está en R Markdown y su versión ejecutada *Practica_2.html*

Contribuciones	Firma
Investigación previa	ARSG
Redacción de las respuestas	ARSG
Desarrollo del código	ARSG
Participación en el video	ARSG